大数据应用过程中的质量控制

浙江财经大学

李金昌

2019.10.18

一、引言

- 数据是最短缺的资源，而大数据是如今最重要的数据存在形式，也是政府统计的重要来源之一。
- 无疑，我们要珍惜和应用大数据。但必须认识到，对使用者来说，大数据并不一定是理想的数据，它同样存在质量问题，并且可能比传统数据更为复杂。因为，大数据存在以下六个不确定性：
- 数据总体的非确定性；
- 数据表现的非标准性；
- 数据含义的非单一性；
- 数据产生的非独立性；
- 数据真伪的难分辨性；
- 数据覆盖的不完整性。

二、什么是大数据质量？

- 对于什么是大数据质量，目前并没有一致的认识，但基于数据的一般特征，其质量应该包括及时性、完整性、准确性和适用性等因素。
- 所以，所谓大数据质量就是基于以上这些因素而衡量的大数据满足使用者需求的程度。
- 大数据质量的特征：
- 1.它不是“无中生有”的质量，而是“有中选用”的质量；
- 2.较好的及时性；局部的完整性；相对的准确性；较差的适用性。

三、大数据质量的主要问题

* 1.与数据使用目的的契合度可能比较差
* 　　大数据通常不是针对特定数据使用目的而产生的，它是现代信息技术应用的副产品。对于数据使用者而言，大数据是自然生成的，具有很大的不确定性，它不像传统数据那样具有很强的目的性和针对性。因此，能否从中选到合适有用的数据，显然是一个无法回避的问题。

* 2.产生系统性误差的可能性更大
* 数据误差的构成是极其复杂的，有些属于客观原因，有的属于主观因素，但大的方面可以分为两类：偶然性误差和系统性误差。相比较而言，系统性误差由于难以甄别、难以测度而更令人头疼。大数据可能由于数据覆盖面不全、社交人群之间的相互影响、个体小数据的倾向性虚假等原因而更容易产生系统性误差，代表性往往难以得到保证。

* 3.数据的可比性问题可能更为突出
* 传统数据由于事先有严格的指标定义、测度标准、获取范围、具体来源、获取方式和衔接调整方法，可以保证数据在时间上、空间上的可比性。但大数据由于其动态可变性、型态复杂性和多变性、含义非标准性与涌现性、语境的差异性、来源的区域性、分类储存的非统一性等原因，使得数据更可能在时间上缺乏连续可比性，在空间上缺乏横向可比性。

* 4.其他相关质量隐患
* 例如：
* 数据的追踪与审核可能受数据的所有权、商业机密与隐私保护等问题而受阻（给数据评估带来困难）；
* 分布式数据的匹配性可能较差（给数据建模带来困难）；
* 数据孤岛问题可能造成逻辑混乱（给数据解释带来困难）。

四、如何控制大数据应用质量

* 1.理论准备
* 改变对数据及其来源的认识；
* 改变对总体、个体、变量等的认识。
* 2.建立完整的质量控制方案
* 事前充分准备（明确数据使用目的，数据源评估与筛选，建立数据分类、测度等标准）；
* 事中同步控制（数据对接，数据审核与修补，数据比较与验证）；
* 事后及时评估（结论的逻辑性检查、合理性评估，使用效果评估，经验与展望）。

* 3.重视对小数据的研究
* 小数据是反映单个人或单个事物特征的数据，多方面的小数据就构成小数据集；
* 无疑，大数据是由大量不断增加的同类小数据所构成的，所以大数据质量取决于小数据质量；
* 要提高大数据应用质量，必须从小数据着手，加强对小数据的研究与评估。只有了解和掌握了小数据的特性，才能真正驾驭和使用好大数据。否则，对大数据的应用就抓不住根。

* 4.对大数据企业进行引导
* 由政府职能部门基于专门研究为大数据公司提供数据定义、数据分类、数据含义、数据储存、数据提供等方面的标准化建议。
* 5.建立与大数据应用有关的法制
* 建立安全、保密与有效使用并重的法制。
* 6.人才培养、培训
* 提高人们使用大数据的技术能力与方法水平。

五、值得注意的几个问题

* 不要在路灯下面找钥匙；（醉汉的故事）
* 不要轻易相信“n=all”；（波士顿颠簸的街道）
* 不要忘了事物的本来规律；（肉牛的命运）
* 不要过分相信纯定量结果；（球探的作用）
* 不要迷信数据量越大越好；（文学文摘与盖勒普）
* 不要陷入“测不准”迷途。（Google流感预测）

谢谢

Thanks